Part A

Main goal of the paper

In short, the main goal of the paper is to evaluate gender inequalities in Wikipedia. Referencing their previous work, the researchers state that women and men are treated in an unequal way as members of the Wikipedia community. It is also stated that men are overrepresented as Wikipedia contributors compared to women – especially white men. In this paper, they aim to find out the extent to which Wikipedia suffers from potential gender bias contentwise – in other words, whether the inequality in the community reflects on the way the two genders are presented. The paper goes in-depth on potential inequalities reflected on biographies, meaning articles written about either male or female personalities. The research on inequalities is divided to five dimensions: notability, topical focus, linguistic bias, structural properties, and meta-data presentation.

One of the main hypotheses in the paper is that the Wikipedia entry points functions as a glass ceiling. The origin of the glass-ceiling effect on Wikipedia is unclear – is it a reflection of our society, or are the members of the Wikipedia community unconsciously enforcing it further by holding different standards for men and women in terms of notability? The aim is to find out how the gender gap affects the entry barrier in Wikipedia, in other words, whether it's easier for a less notable man to make it into Wikipedia compared to a woman with roughly the same level of global notability. As a hypothesis, fewer women are depicted on Wikipedia, but those women can be defined as more globally notable, since the glass-ceiling prevents less notable women, especially locally notable, making it into Wikipedia.

Main methods

As discussed in the paper, notability is a difficult trait to define. The researchers have decided to use two proxy measures to define it – internal and external. The first signaling factor of the global notability of a person is the number of language editions containing an article about them. A person that is depicted in multiple language editions is defined as more relevant than someone whose biography only appears in a few editions. This proxy measure is enabled by the DBpedia dataset that provides a mapping for articles between different language editions. Only the biographies that appear in at least one of the top 20 languages of DBpedia are considered. It is then counted how often these articles show up in other language editions, including all 125 Wikipedia language editions. The second proxy measure used is Google search volume, more specifically the number of times a person's full name is entered in Google. The researchers count the number of countries and the number of months between January 2004 and October 2015 in which the search volume is above a certain threshold. To see whether either of these proxy methods are influenced by gender, the researchers fit binomial regression models that both use gender as the independent variable.

To compare the lexical presentation of the two genders, the researchers focus on topical and linguistic biases. Topical bias is researched by associating the top 200 n-grams for each gender with four topics predicted to be over-represented in articles about women: gender, relationship and family (plus an unrelated category, other). The n-grams are ranked using Pointwise Mutual Information (PMI), which measures the relationship between gender and an n-gram. After ranking the top 200 n-grams for each gender, chi-squire tests are used to compare the portion of topics that are present in both top lists. Linguistic bias is researched by counting the numbers of positive and negative words for each biography. As adjectives are the most abstract class, their extent in a biography is used as a measure of abstraction. The main method is to quantify the tendency to use abstract language as the ratio of adjectives among positive and negative words. This way, we can see the extent to which positive and negative aspects of the biographies of each gender are generalized.

In terms of structural properties, the paper aims to investigate how the visibility and reachability of an article depends on the gender of the person it is about. Using chi-square tests, the relative proportions of attribute presence in the infoboxes of biographies between genders is compared. Since hyperlinks in Wikipedia articles are used to build a network of articles and affect heavily the visibility of articles, they are also analyzed in the paper. The effect of gender in the connectivity of articles and well as the relation between a person's centrality and their gender are investigated. The latter one is done by computing the PageRank – a measure of network centrality – of articles about people, sorting the biographies in descending order according to their PageRank values, and finally, estimating the fraction of biographies about women at different ranks.

Key findings

As predicted, local heroes (people whose biography appears in only 1 language edition) have a high men-to-women ratio, confirming the fact that the entry barrier into Wikipedia is in fact higher for women than it is for men. This might also reflect the gender inequalities present in other media, since Wikipedia editors often rely on secondary information sources. Using a binomial regression model, it is revealed that women on Wikipedia are, on average, 13% more notable than men. The glass-ceiling effect can, however, only be claimed for women born after 1900, since for the women born before that year the results of the regression model are more modest. The external notability proxy based on Google search trends also confirms this theory. Women in Wikipedia are of more interest globally according to search volume statistics, as they are searched in more regions and during more months than men on average.

The differences in n-grams are more significant in the pre-1900 dataset, since bi-grams strongly associated with women born before 1900 relate frequently to topics such as gender, relationships and family, whereas bi-grams associated with men circle around other categories, such as sports and politics. According to the chi-square tests, gender-specific differences are less prominent in articles about people born post-1900. The results on linguistic biases follow the predicted directions. It is shown that abstractness in the form of adjectives is used around 10% more in describing men's positive aspects, whereas almost 2% more in describing women's negative aspects. In assessing the differences in attributes extracted from the biographies' infobox templates, it is suggested that they can mostly be explained by the positional differences of men and women portrayed in Wikipedia. Sports-related attributes, such as position, team and years are more prominent in men's biographies, whereas art-related attributes are more commonly found in women's biographies. Some attribute differences also correlate with the results about topical gender differences - the attribute spouse is more frequent in biographies about women. this is a term related to relationships, which is a topic more commonly discussed in women's biographies.

As the top 30 biographies about men and women - respectively - are sorted by PageRank, it is revealed that the top-ranked women are slightly less central than men, and that their centrality decreases faster with their rank. Considering the top 1000 biographies according to their centrality rank, the fraction of women is less than 20%. Considering only people born before 1900, the number is even lower, and female biographies make up less than 10% of the top 1000 biographies by rank. The results on the hyperlink structure on Wikipedia indicate that there exists a gender bias that favors men. Hyperlinks in men's biographies lead to a biography of another man in over 90% of the cases. Even in women's biographies, over 60% of the hyperlinks lead to men's biographies. This suggests that men's biographies are significantly easier to navigate to.

Further research

In the paper it is mentioned that the dataset used (Wikidata) contains the inferred gender for biographies based on the number of grammatically gendered words. This dataset is also said to report more genders, such as transgender male and transgender female. Due to their small presence, they are left unresearched in this paper, but I would find it interesting to investigate them as well, and compare, for example, how gender bias reflects on the transgender representation in Wikipedia – do the same findings of the gender gap between males and females also apply to transgenders, or are there some key differences? I would use the roughly the same methods as used in this research to look into notability, lexical presentation and structural properties, but switch the data points from men and women to transgender men and transgender women.

Data examination

The DPpedia dataset could be considered as big data, since it contains meta-data for articles in in 125 Wikipedia editions. While the exact number of data points is not specified, it can be inferred that the dataset includes a large number of data points considering the extensive coverage of Wikipedia articles. It is not explicitly mentioned whether the data was collected continuously of at specific times. However, since DBpedia dataset is a structured version of Wikipedia, it is likely to be collected and updated periodically and could thus be labeled as “always on”. This enhances the research's ability to observe long-term trends and avoid missing unexpected events. This also makes the data highly representative, since there is no sampling of larger data: all biographies are considered in the study.

The data is also nonreactive, since the people writing the articles did not know that they would be utilized later on in this research. This means that the writers could not have been adjusting their behavior and intentionally writing more gender-equal biographies, which could have happened if they knew they were being observed. The data utilized by the researchers is not, in my opinion, incomplete. Even though the datasets are not specifically designed for this research, they meet the goal of the study. The purpose is to study gender inequalities present in Wikipedia through biographies, and the dataset is entirely focused on that. It contains all Wikipedia biographies in 125 language editions, as well as inferred gender for those biographies. The data is also entirely available to the researchers with no accessibility problems.

In my opinion, there is no drifting, since the system (Wikipedia) as well as the way people are using it have stayed fairly consistent over the years. I cannot recognize any algorithms in Wikipedia that would be guiding users to write more or less gender-equal biographies, thus the data is not algorithmically confounded. The question about the datasets being dirty is a bit more complex - I'm not sure if there are biographies included that were created or edited by bots. If this is the case, then that would distort the results, since the goal is to study human behavior on Wikipedia. However, without explicit information on the presence of such data, it's challenging to make a definitive judgment. As long as the datasets do not contain highly sensitive or personal information beyond what is typically found in biographies, sensitivity concerns are minimal.

Part D

Main goal of the paper

The main goal of the article is to examine the growth of non-direct work, conflict, and coordination costs in Wikipedia. The article aims to develop tools to characterize these costs and conflicts in order to inform the design of new collaborative knowledge systems. It discusses the exponential growth of Wikipedia and the challenges that arise in terms of conflicts between users, communication costs, and the development of coordination procedures. The article introduces the concept of "indirect work" and "conflict and coordination costs" and proposes quantitative measures for assessing coordination costs. It also presents a characterization model for conflict at the article level and a user conflict model to understand the sources of conflicts in Wikipedia.

Main methods

The researchers analyze the growth and trends of direct work (article editing) and indirect work (discussion, procedure, user coordination, maintenance activity) in Wikipedia. They examine the changing proportions of these activities over time to understand the increase in coordination costs. The researchers also develop a model to characterize conflict at the article level. A CRC (Controversial Revision Count) measure is developed, which refers to the count of article reverts that have been labelled as controversial. A machine learning model is then trained to predict CRC scores from raw page statistics. The model is trained by using articles labelled as “controversial” in their latest revision. The model is also used to weigh metrics based on their utility in predicting CRC scores.

An online survey is used for generalization and validation of the machine learning model. A small set of articles with machine-predicted CRC scores are assessed by Wikipedia administrators, providing a metric of comparison for the model. This provides insight on how well the model performs in identifying conflict even in articles that have not been previously tagged as controversial. As for investigating conflicts at user-level, user and revert statistics – such as the number of users who made at least one revert, and pages with over 50 reverts – are used to understand conflicts. A user conflict model is developed based on revert relationships, meaning measuring the amount of dispute between two users as the number of reverts between them. A Revert Graph is used to visualize user relationships based on revert relationships. The Revert Graph uses force directed layout to simulate relationships between users. This aims to understand the motivations and sources of conflicts among Wikipedia contributors.

Key findings

There is a clear decrease in direct work, referring to edits to article pages. This is primary evidence of increasing coordination costs in Wikipedia. The proportion of direct work has decreased from over 90% of all edits to around 70% from 2001 to 2006. During the same time span, the proportion of indirect work (e.g. conflict resolution and community management) has increased from roughly 2% to 12%. One of the forms of indirect work in Wikipedia is maintenance work, which divides to making reverts and combating vandalism. It is presented in the research that around 7% of the work in Wikipedia goes to reverts, meaning restoring articles fully of partly to their previous versions. The number has been steadily increasing over the years. There is also a small increase (1-2%) in proportion of edits marked as vandalism, referring to a user degrading the quality of an article. The findings suggest that sustained growth in Wikipedia is not solely dependent on the increase in articles and content quality. The ability to coordinate users and manage conflicts is vital for a community where diverse viewpoints exist. These results may have broader implications for other systems involving maintenance activities and multiple viewpoints.

The machine learner discussed before is proven highly effective at predicting CRC from the page metrics. It is found that the number of anonymous edits correlate positively with conflict when made to the article talk page, but decrease conflict when made to the main article page. The model is validated, since the predicted CRC scores correlate with the mean ratings made by users. This suggests that the model is successful in identifying the level of conflict for articles that have no reverts labelled as “controversial”. This highlights the possibility of developing automatic detection and prediction models for complex phenomena like controversies in large-scale social collaborative systems. Building such models can uncover relevant metrics correlated with the phenomena.

The article presents a novel visualization technique for representing conflict between users by mapping revert relationships onto a force-directed graph. This visualization tool helps model conflict in online communities and allows for clustering users into groups based on shared points of view. It can be useful for researchers studying conflict and for users navigating complex relationships in online dispute resolution scenarios. Visualizing conflict through revert relationships is shown to be an effective research tool for making sense of complex relationships between users and quantifying them as either conflict behavior or collaborative behavior.

Overall, the article emphasizes the significance of understanding conflict dynamics, coordination costs, and developing effective conflict management strategies in social collaborative systems like Wikipedia, and provides insights into predicting conflict and visualizing relationships in large-scale online communities.

Further research

If unlimited access to data and compute resources were available, one potential follow-up analysis based on the paper could involve examining the temporal evolution of conflict and coordination costs in Wikipedia. This analysis would involve analyzing how conflict levels and coordination efforts have changed over time and identifying potential patterns or trends. By leveraging the entire history of Wikipedia articles, including data beyond the article's cutoff date, it would be possible to investigate how conflict and coordination have evolved as the platform has grown. This analysis could involve examining various metrics, such as the frequency and intensity of conflicts, the types of conflicts that arise, and the strategies employed to address them. Additionally, with unlimited resources, it would be interesting to conduct analyses focusing on specific domains or language editions of Wikipedia. This would allow for a deeper understanding of how conflict and coordination vary across different contexts and cultural backgrounds.

Data examination

The dataset used in the study is a complete history dump of the English Wikipedia generated on July 2, 2006. It includes over 58 million revisions from more than 4.7 million wiki pages, with approximately 2.4 million article-related entries. This indicates that the dataset used in the analysis covers a vast amount of historical data, including revisions, pages, and article-related entries. Leaning on this, I would classify the dataset as ‘big data’. I would also classify the data as ‘always on’. It is mentioned that the dataset reflects the entire history of all Wikipedia articles up until 2006. Therefore, while the dataset may not be continuously updated in real-time, it can be considered comprehensive and representative of the entire history of Wikipedia.

The article does not explicitly mention the use of sensitive personal data. However, as the study focuses on analyzing the growth, conflict, and coordination in Wikipedia, it is unlikely that sensitive personal information is involved. The dataset used in the analysis is primarily composed of publicly available content and user interactions within the Wikipedia platform – also making the dataset fully accessible to the researchers. Some other limitations of the dataset are discussed in the article. The researchers acknowledge that the data is not designed to meet the research goals precisely, indicating that important data points might be missing. Additionally, the article mentions the limitations of relying on maintenance activities and coordination to sustain growth in Wikipedia, suggesting the need for further research in understanding these dynamics.